Amendments to the Claims:

This listing of claims will replace all prior versions, and listings, of claims in the application:

Listing of Claims:

1. (Currently amended) A method ~~perform~~ performed by a computer system, the method comprising:

extracting, by one or more processors associated with the computer system, a set of uniform resource locators (URLs) from one document or from multiple documents downloaded from a web site;

identifying, by the one or more processors associated with the computer system, a sub-string occurring in the set of URLs as a session identifier, based on at least one of a plurality of rules and based on multiple occurrences of the sub-string occurring in the set of URLs;

generating, by the one or more processors associated with the computer system, a clean set of URLs, derived from the set of URLs, by removing the session identifier;

determining, by the one or more processors associated with the computer system, additional URLs that have already been crawled based on a comparison of a clean set of the additional URLs to the clean set of generated URLs; and

where the clean set of the additional URLs is generated by removing another session identifier, or the identified session identifier, of the additional URLs.

2-3. (Canceled)

4. (Currently amended) The method of claim 1, where the comparison of the clean set of the additional URLs to the clean set of generated URLs comprises:

calculating a first fingerprint value ~~derive~~ derived from the clean set of additional URLs and a second fingerprint value ~~derive~~ derived from the clean set of generated URLs, and where the comparison is based on a comparison of the first fingerprint value with the second fingerprint value.

5. (Currently amended) The method of claim 1, where the at least one of a plurality of rules comprises:

determining that the sub-string does not reference content.

6. (Canceled)

7. (Currently amended) The method of claim 1, where the at least one of a plurality of rules comprises:

determining that the sub-string contains characters consistent with a session identifier.

8. (Currently amended) The method of claim 1, further comprising:

downloading content from the additional URLs when the additional ~~URLS~~ URLs are determined to not already have been crawled.

9. (Previously presented) The method of claim 1, further comprising:

storing information based on the clean set of URLs for use in later determining whether the additional URLs have already been extracted; and

storing the set of URLs, including embedded session identifiers, for use in later accessing the set of URLs.

10. (Currently amended) A method performed by a computer system, the method comprising:

downloading, by a communication interface associated with the computer system, one or more documents from a web site;

extracting, by one or more processors associated with the computer system, a set of uniform resource locators (URLs) from the downloaded one or more documents;

identifying, by the one or more processor associated with the computer system, a ~~sub-sting~~ sub-string occurring in the extracted set of URLs as a session identifier, based on the sub-string having a structure consistent with session identifiers and based on multiple occurrences of the sub-string in the extracted set of URLs;

generating, by the one or more processors associated with the computer system, a clean set of URLs from the extracted set of URLs by removing the identified session identifier;

determining, by the one or more processors associated with the computer system, whether additional URLs have already been crawled based on a comparison of a clean set of the additional URLs to the generated clean set of URLs;

where the clean set of the additional URLs is generated by removing another session identifier, or the identified session identifier, of the additional URLs.

11-12. (Canceled)

13. (Previously presented) The method of claim 10, further comprising:

storing the generated clean set of URLs.

14. (Previously presented) The method of claim 13, further comprising:

adding a generated session identifier to each of the generated clean set of URLs.

15. (Currently amended) A device comprising:

a memory to store instructions; and

a processor to execute the instructions to ~~implement~~:

~~at least one fetch bot to~~ download content on a network from a single web site;

extract URLs from the downloaded content;

identify a sub-string as a session identifier from the URLs extracted from the downloaded content based on at least one of a plurality <u>of</u> rules and based on multiple occurrences of the sub-string in the extracted URLs;

create a clean set of URLs by removing the session identifier from the extracted URLs;

store the clean set of URLs; and

determine whether additional URLs have already been crawled based on a comparison of a clean set of the additional URLs to the created clean set of URLs, where the clean set of the additional URLs is generated by removing another session identifier, or the identified session identifier, from the additional URLs.

16. (Previously presented) The device of claim 15, where the processor is further to identify the sub-string as a session identifier based on locating characters consistent with a session identifier in the URLs extracted from the downloaded content.

17. (Previously presented) The device of claim 15, further comprising:

a database to store the downloaded content.

18. (Previously presented) The device of claim 15, where the processor is further to determine whether the additional URLs have previously been stored by comparing the clean set of the additional URLs to the stored clean set of URLs.

19. (Previously presented) The device of claim 15, where the session identifier includes characters from the extracted URLs that do not reference content.

20. (Currently amended) A system comprising:

one or more server devices comprising one or more processors to:

download one or more documents from a web site;

extract a set of uniform resource locators (URLs) from the one or more documents downloaded from the website;

identify a ~~sub-sting~~ sub-string occurring in the set of URLs as a session identifier, based on the sub-string including characters ~~that are structured~~ consistent with session identifiers and based on multiple occurrences of the sub-string in the set of URLs;

generate a clean set of URLs from the set of URLs by removing the identified sub-string; and

determine whether additional URLs have already been crawled based on a comparison of a clean set of the additional URLs to the generated clean set of URLs[[;]], where the clean set of the additional URLs [[are]] is generated by removing session identifier, or the identified session identifier, of the additional URLs

21-23. (Canceled)

24. (Previously presented) The system of claim 20, where the one or more processors are further to:

add a generated session identifier to each URL in the generated clean set of URLs.

25. (Currently amended) One or more memory devices that include programming instructions executed by one or more processors; where the programming instructions causes the one or more processors to:

extract a set of uniform resource locators (URLs) from one document or from multiple ~~document~~ documents associated with a single web host;

identify, in the set of URLs, a sub-string as a session identifier based on the sub-string having at least a specified measure of randomness and based on multiple occurrences the sub-string in the extracted set of URLs; [[and]]

generate a clean set of URLs from the extracted set of URLs by removing the identified session identifier; and

determine, by the one or more processors associated with the computer system, additional URLs have already been crawled based on a comparison of a clean set of the additional URLs to the clean set of URLs[[;]], ~~wherein~~ where the clean set of the additional URLs [[are]] is generated by removing another session identifier, or the identifier session identifier, of the additional URLs.

26-28. (Canceled)

29. (Currently amended) The one or more memory devices of claim 25, ~~further causes~~ where the programming instructions further cause the one or more processors to:

add a generated session identifier to URLs in the clean set of URLs when the URLs are to be used to access a web document.

30. (Currently amended) The method of claim 1, where the at least one of the plurality of rules comprises:

determining that the sub-string exhibits at least a specified measure of randomness.

31. (Currently amended) The method of claim 10, where identifying the ~~sub-sting~~ sub-string occurring in the extracted set of URLs as a session identifier includes identifying the sub-string as having at least a specified measure of randomness.

32. (Currently amended) The device of claim 15, where the processor is further to execute the instructions to:

identify the session identifier from the extracted URLs based on identifying that the sub-string exhibits at least a specified measure of randomness.

33. (Currently amended) The system of claim 20, where the one or more processors are further to:

identify the ~~sub-sting~~ sub-string occurring in the set of URLs as a session identifier based on the sub-string having at least a specified measure of randomness.